NPTEL

NPTEL ONLINE COURSE

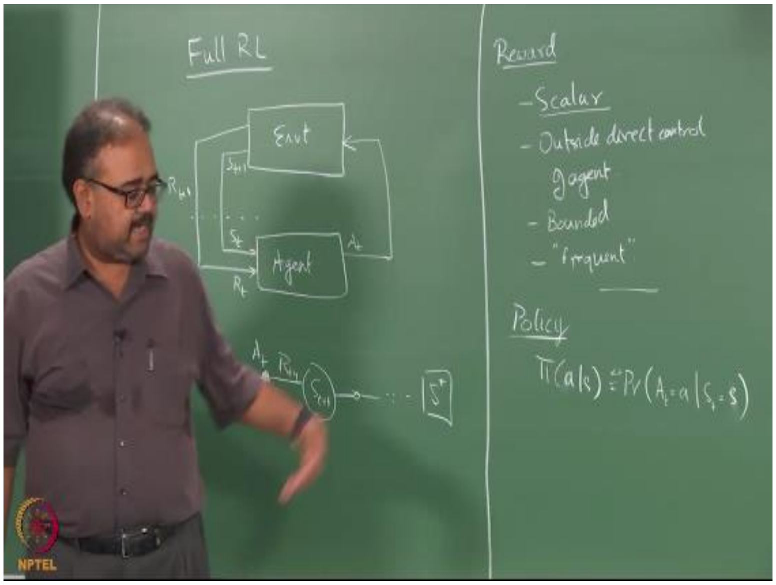
REINFORCEMENT LEARNING

Full RL Introduction

Prof. Balaraman Ravindran
Department of Computer Science and Engineering
Indian Institute of Technology Madras.

So far we have been looking at bandit algorithms right or the immediate reinforcement learning and so what we are going to start from today is looking at the full reinforcement ionic problem right.

(Refer Slide Time: 00:40)


Full RL
Envt
S_{t+1}
R_{t+1}
S_t
Agent
A_t
R_t
A_t R_{t+1} S_{t+1} ... S^+
Reward
- Scalar
- Outside direct control of agent
- Bounded
- "frequent"
Policy
$\pi(a|s) \doteq Pr(A_t = a | S_t = s)$
NPTEL

So we know earlier already about the full RL problem right so what are the complications that we are going to insert here States we already looked at states in contextual mandates right temporal dependents right the sequential nature of the problem is something that we are going to introduce here so what we are going to look at is that we will assume that there is a there is an environment okay and the agent interacts with the environment at any point the agent senses the state the environment is in okay.

Then in response to that the agent takes a action which is then applied to the environment okay so the agent sensors the state the environment is in SST right and it takes an action that gets applied to the environment and as a result of applying this action environment changes to a state St + 1 which then gets fed back into the agent has the current state St and then you keep doing this in a loop right so this could go on indefinitely then all there could be a well designated stopping state right if you reach the state you stop the trends okay a group of states I mean you can always this is the strict right.

So when I say that is a starting state and the stopping state it could very well be a group of starting states and the group of stopping states I can always define a dummy state to which all the states go to right so I can always say that is a single starting state in a single stopping state right so that if a reach of exposition so I will keep saying that makes sense and even if there are multiple states where you will stop I can insert one dummy state and say that automatically as soon as you reach one of these states you will make a transition to the dummy state and you will stop them okay.

So they will always be a one-stop state one starts taking right so that might happen and apart from this you also get a remote signal from the environment okay this is like as mentioning earlier this is a convenience a modeling convenience that we put the reward in the environment okay that is because we want to put the reward in a place which is outside the agents direct control right agent should not this arbitrarily say okay I okay I am getting a reward of five now right.

So it should get this reward as a result of the behavior in the world okay it cannot just arbitrarily set its own rewards okay so that is the reason we put the reward in the environment okay so that is outside in reality right this will be a very much more complex picture right because if you think about it I mean I have been repeatedly saying that all biological agents of reinforcement learning agents right humans are reinforcement learning agents so do you have a separate reward conduit right.

We have a reward sensing mechanism somewhere right so you only have what you sense from the world right so you look around you see whatever is whatever inputs that are coming in whether you are in you could being interesting food you could be looking at the outside world right somebody might be touching you on the skin whatever ray so you are taking in all these complex sensory inputs and other kinds of actual physical ingestible inputs okay and then you are converting them into rewards right so that is how the rewards happen right.

So if you are hungry and you eat okay that is that is rewarding right but then you and find out that you have I do not know eaten something that is gone stale right let us not rewarding it is in fact opposite of it right and if you have not hungry and somebody just gives you something to eat you just eat for the sake of you know I give me company and then you go and eat some really bad food and things like that so there is no immediately water part from the company part of it but the act of eating itself does not give you any immediate reward but probably we have a consequence in the long run Yeah right so go eat some right foot just because your friend asked you to eat will have consequences in the in the long run right .

So rewards are if he thinks right I mean so we are going to make it very concrete right I am going to say that is this some mechanism in the environment that gives you a reward but if you really think about it these rewards are or everywhere right in fact the last few years so Andy Balto was asked going around asking this question where do robots come from right so when was he here 2009 yeah so I hosted him here in 2009 and one of the talks I mean at that time the talk he was going around giving everywhere was flagged or watts come from okay so it was it is really an unforgettable thing yeah.

So also this is not very clear so take it with a pinch of salt when I say that the environment is going to give you a reward signal like the robots could come from anywhere right in fact it is not even clear what is the agent what is the environment so you have a lot of Robotics guys here right suppose I am having a robot learning problem right so what would be the agent what would be the environment the robot is agent really okay let us say I give you a navigation problem or a reaching problem let us say I am trying to train a robot to reach out and pick an object on the table.

In the environment is the workspace where were they armed no but what about the angle of the arm we sit in the environment or is it part of the agent so let us see where is agent then so the arm is not part of the agent the one controlling the arm is the Agent so it becomes tricky various we are actually think about it is not like okay as a robot the robot is a agent everything outside the robot is environment right so it is not that boundaries are not drawn as neatly around the physical entities right suppose let us consider forget about robots let us say you are driving a car okay so who is the agent what is the agent and what is the environment there.

Right I mean always the right answer is depends right but what does it depend on right in this case the agent could be the car wait the environment could be the road right so as a human being sitting and driving a car you can look at okay what is the velocity which the car is moving okay and what is the direction which the car is moving about pedestrians jumping in front of me right and all those things right so this could be your and it just could be your environment right and the agent itself is the car you know I am controlling the car on the other hand the environment could include the car right and the agent could be the human sitting there and you can think of okay.

How much how much am I pushing my foot down on the pedal which petals should I be keeping my feet on first case right and then should I how much how about what talk should I apply to the steering wheel right where should I be turning my head to look at right would I be looking at the mirror should I look at front right or should I be looking down at my phone right so all of these things decisions that you have to make right while you are driving right so there is this interface now the agent environment interface becomes a human when the car right or it could be even more right he can go further back it can even go further back into the human right thinking about okay which okay I am talking about the motor cortex right.

So which neurons in my motor cortex should for now so that my arm switches enough so that I turn the wheel this much right so which neurons should fire in my motor I mean the autonomous nervous system so that my leg actually pushes down with some kind of pressure right so I could I could think of depending on the kind of problem that I am solving and I could think of drawing this agent environment boundary anywhere right and then not arbitrarily anywhere but you can think of multiple places where the same set of interacting entities are there right but you have very different kinds of modeling that you could do.

Like it is not very clear when I say there is an agent there is an environment do not expect people to come and give you okay this is the agent here is environment now come up with a reinforcement learning algorithm that will solve this problem right so typically you will have to think about this figure out what is a right abstraction that you would need right before you go so I start a solution for this right so this is something which you have to keep in mind nobody is giving this to you a priori right so having said all of this or one other way of thinking about how this interaction proceeds right.

So model all the caveats I have been giving to you there is an agent okay it senses the state of the environment it takes an action okay and then that causes a change in the state of the environment right so having said all of this you can think of the interactions as proceeding like this so I have St right I sense that I take action 80 okay that causes new state St + 1 under reward so this is a slight quark of the book itself right where the reward corresponding to the state and action at time T is labeled as RT + 1 okay so this is a quirk in the book and the reason they say is typically the next state and the next reward or determined at the same time.

And therefore it is a makes more sense to label both of them with T +1 okay so one thing here in the Bandit case I was making a distinction between N and D okay I am going back to using T here but the T could be discrete time as well in fact for most of the course RT is going to be discreet okay in fact T did not even be time T did not even be time okay so T is steps so what are you talking about here typically these are my decision apex right when I say T = 1 that means I am taking my first decision you say T equal to 2 that means I am taking my second decision right this is typically okay.

Sometimes I am also interested in cases where I do not take any decisions but the state keeps changing over time but for most parts when I say when advanced / 1 that is typically at the point where I'm required to take the next decision okay so later we will actually look at a still a discrete-time case but where decisions need not happen at every time tick so we look at that case as well later but not immediately okay so most part of it assume that T is a decision tick ok not a time tick when T goes from T to T +1 that means have taken other decisions the decisions could not need not necessarily take the same time.

For example consider the game of chess right so each move that you make I mean so initially you will be making moves very rapidly right because you are playing from a standard opening play or something and later on you actually have to study their board position and then you have to think about it and make moves right but I still count each everything as a single move right so you do not say the first few moves were all count as one fourth of a movie each or anything because you did it very fast so just exactly like that so each move itself could take different amounts of time right.

I am not worried about it and the intervals at which you are asked to make decisions might also be non-uniform I am not worried about it okay this is the basic setting I am having later when we talk about hierarchical RL will start worrying about the duration of decisions if I take one action now how long will it take for that action to complete and those kinds of things we worry about later right now we are assuming that every action takes the same time to complete so time with encodes here okay.

So that is basically hit this case keeps going let us oh will denote by S^{+} that distinguished terminal state right so this just keeps going until I S^{+} or it just keeps going in definitely if I there is no distinguished terminal strength okay any questions of a setup is clear right so what we are going to do okay next thing I want to talk about is the reward itself so I am going to be very religiously following chapter 3 in the book right and I did ask people to read ahead to Chapter three how many of you actually read chapter three, four my god you did you put your hand up.

Okay then not content so 4 people yeah sorry you are busy doing the assignments is it he will release the assignment from chapter 3 as well so from now onwards what we will do is before I start the chapter will release the assignments will it help letting the same of course not the programming assignment yes so when is when is the second return assignment due Friday is it so if you release some questions today they will be able to finish it before Friday night I only promise I will not release it on Thursday.

Then I promise I will not release questions on Thursday right I did not say anything ever Wednesday a world is not giving any opinion you were just thinking about okay they have not do anything to him this I am the one who will get beaten up okay so it will extend the deadline to Monday and then give them additional assignments today yes so the people who read the head so what does the chapter say about reward functions this is anything but word shear anything else about the robot so classically reinforcement learning assumes that the reward is a scalar quantity alright so the rewards are so the words are scalar and it is outside the direct control of agent.

That is why I told you we put it in the environment yeah we will come to that right so what does it mean that agent cannot arbitrarily set the reward it has to do it through the interaction with the environment okay so that is essentially the two main things that we need so the more important thing here is the assumption that the rewards are scalar right so what does it mean it means is just a number so it might look that it is a pretty weak model for the rewards but if you remember we talked about a wide variety of applications of reinforcement learning in the first lecture and then all of those applications that we went through we were actually looking at scalar robots.

Right so it does turn out that variety of settings where you can use the scale or what and get away with it right but there are in cases where you might have to make some kind of calls about trade-offs between how much importance you want to give to one event versus another so for example let us again take the reaching tasks that we had right so the so when I want to reach out to an already set the rewards here once a graph it I will give a reward right I can give you some more like can give you a reward of 10 or something when I graph the yeah but then I need to optima need to get there right.

So I what I am trying to do is optimize my reward function if you remember the whole problem in the reinforcement learning is too well I see something if you mentioned in the very beginning the problem in RL is to behave in such a way so that you get as much as reward as possible in the long run right so if I do not give you any reward function I will just be flailing around right I could be happily moving waving my hand around not doing anything in regard to grasping the object right.

So I need to give some kind of a reward either for grasping the object or how else would you do this negative reward for every time step I keep flailing around right it is like you are in constant pain right and the pain will stop once you hold this right because the episode ends and you stop getting minus once then right so like somebody keeps poking you with a pine all the time keep go and you reach this is how we try yeah anyway any of any animal that draws a vehicle right you keep pushing it in the side until it actually reaches the destination right.

I mean and then humans are very cruel man I mean anyway so that is how we do it right so but then instead also set sufficient so you do not really want to strain your hand right so I mean there are joint limits right so you really do not want to come up with a crisp grasp that looks like this right this also valid I have got the phone right so you do not want to do that I want to give some kind of you know additional pain right if I do things like this right so if in fact for me it actually was painful right but so you want to do this communicate this to the boat that is how much additional pain do you give it.

So grabbing this thing let us say gives you a reward of 10 right and doing that gives you a reward of say-100 so what is going to happen they are going to be very cautious right in trying to grab it right but if grabbing this gives you the word of 100 right I am doing that gives you a reward of -0.1 you would still be cautious because I am being optimal right so 99.1 miss was than 99.5 so I will still be cautious about avoiding those but I will not go to two significant extent to do that because I am still getting my -1's right so the longer I take to get to the gate to a grasp right the more -1 I am going to get right and the better form the grass place the fewer -1's I am going to get.

So there is some trade-offs I will form at some point right but where the trade-off quite lies depends on the magnitude of the reverse I am going to give you as well so if I am going to give you a -1000 for doing something like this right then I will be more careful right so these are all trade-offs right when you design the problem but so how do you set this reward functions well understood problems like grasping or reaching or robotics in general right you can look at the problem that you are trying to solve and try to figure out okay which are undesirable configurations okay,

So very expensive robot so I do not want to break it so let me make sure that all these other things have a highly high negative values right Andorra versus a cheap robot or what I really want to do is get there as soon as I can okay do not mind throwing 300 robot set it because each robot costs only 1rs right then I can just not worry about crashing and burning a few along the way right so it depends on the domain right so there is still a significant trade off that I will have to think about right so even though I am saying it is a scalar quantity there are instances where I will have to kind of put multiple outcomes right multiple outcomes on the same scale it makes it a little tricky just like in this case I mean that is an outcome associated with success which is grasping it and that is outcome associated with well not really failures but outcomes associated with undesirable events along the way.

So I have to put all of these on the same scale and I have to do it might have been more convenient if you could have thought of okay 104 that k- 5 for this but these are just things that you need to avoid know if there is some other way I can communicate these things to robot then that might be a better mechanism it turns out it is hard to come up with the appropriate optimization set up for doing this so the scalar robot thing A is pretty powerful and B kind of gels in some sense with how we also do decision-making at the end of the day you are very good at judging trade-offs right.

So whenever as soon as you start thinking of doing these kinds of trade-offs in some sense you are putting these on the same scale right the putting them on the same scale because they only then you can think about trade-offs so in that way also it seems to be natural and the third thing is having a single scalar value makes it easier for us to optimize the solution right we can always get a good optimal solution this is for various reasons so scalar values are fine and so I think that is one of the homework questions which ask you to think about scalar values a little bit more I mean think and reason about it.

And so what else do we need the reward to do well mathematically we need the rewards to be bounded we are talking about biological systems is not really a criterion and because I mean it has to be physical right so you cannot have an unbounded neural spike in your brain for corresponding to a robot right so you may just fry your brain right so but when talking about mathematical a certain settings you really want through what to be bounded right so basically that is it so one other thing later on we would like the robot to be frequent I will explain this later like once we start looking at learning algorithms and system but this essentially means that it should not be that I get rewards say once in a million time steps.

There is very little very limited literature in this right so the one alternative which people have explored is to look at vector-valued rewards where each component of the reward corresponds to a different thing that you want to achieve and these are called typically multi-criteria RL right there are very few papers or maybe five or six papers out there which talk about multi criteria reinforcement learning in fact there is a corresponding optimization thing also called multi-criteria optimization right.

So where people talk about different notions of optimality event so what exactly is optimal rate so optimal is where the objective function you reach a minimum of objective function but now if you went look at it in multiple dimensions so you may end up reaching some kind of saddle points also right so what exactly is the Optima in this case so people talk about those kinds of characterization.

So there is more work in multi-agent settings in looking at more richer reward functions than in single agent RL right yeah very little out there and it will be interesting to come up with some kind of qualitative feedback as well instead of looking at this kind of a quantitative feedback and if you like that kind of work there are there is some kind of work in fuzzy RL where the feedback signal is a fussy signal not much again very little and one more things I want to mention but all of those are scalar I mean there are all kinds of work on like instead of looking at the reward but look at the perception of reward and so on so forth.

But all of these are still scalar quantities and they are not vector valued quantities but it is a much h richer notion of feedback then what you what is typically a reward functions so what we mean by frequently said you should not be getting rewards like once in a million time steps or something then you will never learn anything right you are trying to optimize rewards and regardless of what you do for a million time steps you get nothing is too long a horizon for you to learn anything over right so typically you expect more reward to come to you along the way okay.

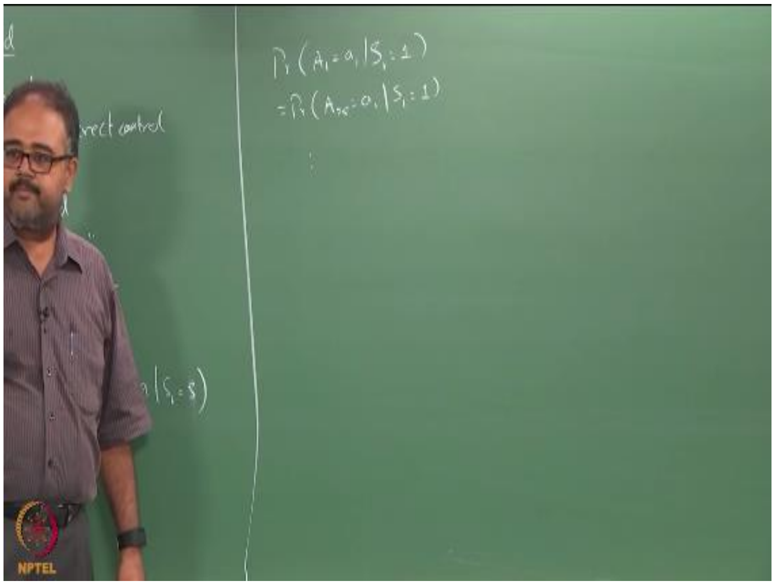
So that is what robot so policy so what is it so I am just talking about different components of a reinforcement learning system okay so we talked about states and actions right basically the agent and the environment now you're talking about the robot okay the next thing you want to talk about is the policy so what is the policy no 0 is non infirmity of right I really cannot change my unless I have gotten a -1 so where I do not know if zero is good or bad I mean I really cannot change anything based on 0 but that is a good point yeah.

So when we say frequent rewards I do mean frequent non zero rewards right so we know what the policy is right all of us know what policies are so what is the policy so I am going to denote a policy as this as a conditional thing from now on so what is this is essentially the probability that there is a probability that at time T the agent will take action eighth given that at time T the state is s.

So ideally I should have a T there right so π T(A, S) is a probability that at time T I will take action A given that at time T the state was S okay so normally for most of the material that we cover right I will be looking at what are known as stationary policies so what is the stationary policy the stationary policy is something which is essentially so regardless of the value of T I will be using the same probability so the probability that let us say my state is 1 right probability that when give me some other symbol to use a state let us say my status l okay I do not think I used L for anything so far have a we have in the proofs you use L or is it yeah let us say okay fine let us use it one then people can all of you are smart enough to figure out this right.

So let us say state is 1 right so I do not care whether it is one that time one whether s1 is equal to 1 or S 25 is equal to 1 the probability that I will take action A1 will be the same right the probability I will take action A2 will be the same so essentially what I am saying is.

(Refer Slide Time: 32:09)



Probability that A1 equal to a1 give S1 = 1 = probability that something like this okay right so this is what I mean by stationary that is what I am saying here so if I want to explicitly talk about non-stationary policies okay I will use the π T notation okay if I am talking about stationary policies I will just use the π notation.

Okay so what is my policy the what is a goal in trying to come up with a policy see one even go there it I am just defining things right a non-stationary policy is just a collection of functions collection of probability distributions one for each time step at a stationary policy is a single well it is actually also a set of probability distributions one for each state right a non stationary policy is a collection of such sets one for each time step as well as one for each state that is basically it how you derive how we derive anon-stationary policy or how we learn it from data and other things they are independent questions I am just talking about it is simple definition of the quantity when you learn your goal should be a stationary policy right the end goal can be a stationary policy where you have a stationary there is no subsequent we will come to that right so there are different time scales at which you can learn right.

So I can have a stationary policy run the full episode using the stationary policy and then take the rewards I get and then use that update my parameters and come up with a new stationary policy right so these times I am talking about the T here is essentially the index over a single interaction okay I am not talking about multiple interactions over the single interaction right suppose I start from some s0 I come and end at s^+ over that interaction my policy stays fixed that seems to be a reasonable assumption right.

So you could think of having states learning a sequence of stationary policies so what is my goal in learning a policy here is to learn something that will maximize the reward that I get right but not individually but something that will maximize it in the longer okay so what do I mean by that suppose I am riding a cycle right so suppose you leave your hands and say they look at me or something like that might give you a temporary kick you know I feel really good about myself but then you met the next minute you might ram into another vehicle or a tree or a building whatever right and get injured right.

So that is something that gives you a depends but hopefully there is something that will deter you from doing crazy things on a cycle the next time around right so the short-term reward was just feeling good about you for one minute and the long term penalty was getting into a cast or whatever right so these kinds of things are odd or even simple things like in chess you know the first thing kids they try to capture the Queen and whole bunch of other things and then as a result end up losing the game right so this can something so you do not look for short-term rewards you should look for something that maximizes your rewards in the long term.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in